ПЕРСОНАЛЬНАЯ ЦИФРОВАЯ БИБЛИОТЕКА LIBMETA КАК СРЕДА ИНТЕГРАЦИИ СВЯЗАННЫХ ОТКРЫТЫХ ДАННЫХ

Атаева О.М., Серебряков В.А.

Москва, ВЦ РАН
oli@ultimeta.ru, serebr@ultimeta.ru

***Аннотация**: В статье описывается семантическая электронная библиотека Libmeta, ресурсы которой могут быть обогащены за счет использования данных из источников, расположенных в LOD. Связывание происходит посредством онтологии предметной области, которая задается пользователем и определяет его область интереса. Рассматриваются проблемы интеграции ресурсов библиотеки в LOD и создания поисковых запросов по источникам данных, а также обсуждается использование спецификаций и технологий из стека LOD в рамках одной системы.*

***Ключевые слова**: цифровые библиотеки, semantic web, онтологии, linked open data.*

1. Введение

Последнее десятилетие наблюдается бурное развитие технологий Semantic Web и активное развитие сообщества, поддерживающего Linked Open Data (LOD). Основная идея LOD заключается в решении задач интеграции данных, представленных в сети, для чего предлагается представить информацию в формализованном виде, что делает ее доступной для машинной обработки.

Развитие технологий Semantic Web и популярность идеи LOD оказали влияние и на электронные библиотеки, которые трансформируются и превращаются в центры данных, вокруг которых формируется сообщество заинтересованных экспертов и пользователей, принимающих активное участие в их развитии. При консорциуме W3C была создана рабочая группа под названием Linked Library Data, которая выработала рекомендации по связыванию библиографических данных с использованием стандартных семантических технологий RDF, SPARQL, OWL. Появление семантических технологий вызывает необходимость разрабатывать новые подходы к созданию электронных библиотек и расширяет возможности их использования.

2. Эволюция библиотек

Развитие информационных технологий в 20 веке и их использование в библиотеках привело к появлению нового типа библиотек

2.1. Электронные библиотеки

Электронные библиотеки представляют собой набор документо-подобных ресурсов и их библиографии, а также сопутствующих услуг для их хранения и поиска. При этом не выделяются другие виды важных объектов, например, персоналии, организации и т.п. Встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Даже идентифицировав персону, как правило, нет возможности получить документы, связанные только с ней. Это обусловлено тем, что метаданные рассматриваются как нечто, связанное только с документом.

2.2. Цифровые библиотеки

Цифровые библиотеки представляют собой информационные системы, которые обеспечивают задачи коллекционирования, хранения и навигации по разнообразным электронным документам, хранящимся как в самой системе так и доступных по сети.

2.3. Семантические цифровые библиотеки

Использование семантических технологий значительно расширяет функциональность библиотек, данные лучше структурированы, выделены связи между ними, улучшается поиск, появляется возможность интегрировать данные различных типов: персоны, ресурсы, пользователи. Обеспечивается интероперабельность с другими системами, не обязательно являющимися библиотеками, так как основной задачей семантических технологий остается предоставление метаданных в машиночитаемом формате. Онтологии играют основную роль для решения задач, вызванных структурными различиями существующих систем и семантическими различиями стандартов метаданных.

Выделяют следующие виды онтологий используемых в семантических электронных библиотеках [1]: а) библиографические – для хранения информации о ресурсах библиотеки, б) пользовательские – для описания профилей пользователей библиотеки, включая их интересы в определенных ресурсах и другую информацию, в) структурные – для описания содержания ресурса.

2.4. Персональные семантические цифровые библиотеки

Мы выделяем персональные семантические цифровые библиотеки, наполнение которых индивидуально для каждого пользователя системы и выполняется в полуавтоматическом режиме из разнородных источников данных, интегрированных в облако LOD. Будем далее для краткости называть их персональными открытыми цифровыми библиотеками или ПОЦБ.Типы информационных ресурсов и их структура определяются пользователем,

исходя из своих интересов, то есть пользователь описывает интересующую его предметную область, определяя тематическое наполнение библиотеки.

В работе [2] представлена общая схема системы, выделены ее основные модули и дана характеристика каждого из них. Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности автоматизированного извлечения интересующей пользователя информации по определенной предметной области.

Представление ресурсов библиотеки в виде связанных данных расширяет функциональность семантических цифровых библиотек, давая возможность:

- включения дополнительных элементов описания данных информационных ресурсов,
- полного или частичного обновления данных из источников,
- использовать интерфейсы для создания запросов к интегрированным в LOD источникам данных на основе SPARQL,
- включения в описания ресурсов других типов информации.

Одна из задач, которая решается в ПОЦБ, - это реализация интеграции набора данных в пространство LOD с использованием онтологии предметной области информационных ресурсов, т.е. автоматизированное обнаружение новых наборов данных и по возможности установка и поддержка связей с элементами данных из этих наборов данных с уже имеющимися ресурсами в репозитории библиотеки, обеспечивая одновременно рекомендуемую проектом LOD функциональность в рамках одной системы.

3. Источники данных

Мы подразделяем источники данных на два типа: внешние и внутренние. Внешними мы называем те источники, которые интегрированы в LOD, и данные которых представлены в RDF и доступны нам с использованием SPARQL. Для своих практических целей мы использовали такие известные источники в LOD, как DBpedia [4], Europeana [5]. Внутренние источники могут представлять собой любой другой тип источника данных, который не интегрирован в LOD. На практике в качестве внутренних источников мы использовали другие библиотеки, которые предоставляли доступ к своим данным по протоколу OAI-PMH.

3.1. Внешние источники

Данные из источников LOD хорошо структурированы и обычно доступны через SPARQL точку доступа для поисковых запросов. Так как одним из принципов LOD является использование URI, по которым можно получить по

HTTP информацию в стандартном формате, то для доступа к информации определенного ресурса пользователь может использовать только этот URI.

Основной задачей подсистемы подключения внешних источников, является создание и поддержка отображения онтологии предметной области на схему источника данных, посредством которого пользователь получит возможность автоматического мониторинга для последующего связывания имеющихся данных в системе с новыми данными по определенным запросам в терминах своей онтологии. При этом в системе при импорте может сохраняться лишь внешний URI ресурса.

3.2. Внутренние источники

Несмотря на активное развитие LOD, нельзя игнорировать источники данных, которые в него еще не интегрированы и при этом содержат огромный объем полезных данных. В библиотечной среде для обмена метаданными широко используется протокол OAI-PMH. Основным его недостатком является то, что для доступа к информации о ресурсе нужно обладать специальными знаниями о протоколе, при этом знание URI, который используется в таком источнике, не сильно облегчает поиск этих данных. При импорте данных по этому протоколу мы внутри нашей системы решаем задачу формального предоставления и интеграции этих данных в соответствии с принципами LOD, при этом сохраняя информацию о первоначальном источнике.

В работе [3] предлагается улучшенная версия этого протокола, которая является развитием протокола в сторону поддержки связанных данных.

4. Функциональность ПОЦБ

К основной функциональности системы реализующей ПОЦБ относятся

- Функции атрибутного поиска;
- Функция выделения неявных связей между ресурсами по их описаниям;
- Функции работы с коллекциями;
- Создание/Просмотр/Редактирование/Объединение/Вложенные коллекции;
- Функция отображения онтологии ИД;
- Функция детализации обеспечивает преобразование в подзапросы, соответствующих различным ИД;
- Функция для выполнения запросов и обработки результатов и предоставления окончательного результата пользователю;
- Функция автоматического мониторинга ИД на наличие новых/измененных данных;

- Создание словарей, классификаторов, тезаурусов;
- Редактирование элементов;
- Поддержка («гибкой») классификации ресурсов;
- Поддержка настройки уровней доступа к различным ветвям тезауруса.

Исходя из определения источников данных ПОЦБ и перечня функций системы можно выделить «внутренние» функции, т.е. те, которые оперируют данными в рамках системы и интегрируют данные из «внутренних» источников и фактически определяют обычную семантическую библиотеку. «Внешние» функции обеспечивают подключение и извлечение данных из LOD и позволяют задать тематическое наполнение библиотеки, таким образом задавая фактически определение ПОЦБ.

5. Онтология ПОЦБ

Онтология ПОЦБ, разработана в общем виде без привязки к конкретным методам и способам реализации семантических цифровых библиотек.

Задача построения онтологии сводится к задаче построения онтологии информационной системы на базе онтологии BWW [6]. Обычно при построении информационных систем на первом этапе выделяют общие понятия, которые не зависят от конкретной предметной области. Далее вводятся определения, характерные для конкретной предметной области, которые соединяются с общими понятиями бинарными отношениями.

Фактически общая онтология ПОЦБ состоит из двух онтологий:

1) Онтология СЭБ, построенная на основе онтологии информационных систем, включающая в себя основные понятия, необходимые для обеспечения основной функциональности библиотеки, такие как ресурс, пользователь, коллекция, словарь, классификатор, запрос, источник и т.д.

2) Онтология и тезаурус предметной области, для которой пользователь определяет ее понятия, их тип, структуру, совокупность словарей и классификаторов, которые представляют тезаурус предметной области, который обеспечивает доступ неквалифицированных пользователей, решающих задачи поиска информации, к знаниям предметной области в разных источниках. Эта онтология позволяет:

- выработать и зафиксировать общее понимание области знания;
- представить знания в виде, удобном для их обработки автоматизированными подсистемами, обеспечить возможность получения и накопления новых знаний, а также возможность многократного использования знаний

Тезаурус же обеспечивает терминологическую поддержку и помогает пользователям сформулировать запрос к системе, в том числе, подобрать

правильные ключевые слова для описания искомого результата, имеющихся данных и контекстной информации.

Тезаурус необходим для навигации и для автоматического уточнения и расширения запроса, введенного пользователем, посредством использования зафиксированных в тезаурусе связей между терминами. Например, в качестве предметной области может рассматриваться онтология из работы [7] со всем набором словарей и классификаторов.

6. Поиск по источникам данных

Поисковые системы, ориентированные на источники, интегрированные в LOD, такие как Sig.ma, Falcons, и SWSE, обеспечивают поиск на основе ключевых слов, ориентированный на использование той же парадигмы, что и существующие лидеры рынка, такие как Google и Yahoo. Пользователю предоставляется окно поиска, в котором он может ввести ключевые слова, связанные с предметом или темой, в которых он заинтересован, и приложение возвращает список результатов, которые могут (или нет) иметь отношение к запросу. Фактически это поиск по вхождению слова в любой элемент описания. Поиск же данных в источниках предполагает, что пользователь знает структуру данных

В работе [9] представлена система поиска LOQUS в репозиториях LOD на основе высокоуровневой онтологии, на которую отображается схема подключаемого источника данных (ИД). Эта онтология составлена на основе высокоуровневой онтологии, которая содержит наиболее общие и самые абстрактные концепты, имеет исчерпывающую иерархию фундаментальных понятий (около 1 тыс.), а также набор аксиом (примерно 4 тыс.), определяющих эти понятия. Каждому концепту определен идентификатор или обобщающее понятие из LOD. Онтология, так же, как и в нашем подходе, используется для трансляции SPARQL запросов пользователей в интегрированные ИД. Но недостаточный уровень концептуализации понятий не позволяет в достаточной мере сконцентрироваться на определенной предметной области.

С другой стороны задача автоматизированного поиска релевантных источников данных осложняется тем, что чаще всего информация о связях между ними проставляется в основном на уровне данных с помощью связей sameAs, seeAlso. Даже простой анализ связей sameAs, seeAlso на уровне найденных данных позволит выявить эквивалентные классы, ранее не определенные связи между разными источниками или новые источники. Описание связей на уровне схем затем можно использовать при формировании запросов к источникам данных.

Связи между источниками на уровне схем описываются гораздо реже. В последнее время эта задача решается с введением и активным

распространением спецификации VOID [8] для описания источников RDF данных, в которой предоставляется информация о связанных источниках данных. VOID описание содержит информацию об используемых словарях, статистическую информацию, сколько ресурсов того или иного типа или значений определенных свойств используются во множестве. При создании словаря VOID была сведена к минимуму необходимость создания новых свойств и классов, путем использования существующих словарей. Например, для описания статистической информации используется словарь SCOVO. На основе этой информации можно делать вывод о релевантности источника тому или иному запросу или предметной области.

7. Текущее состояние работ

В рамках создания первой версии ПОЦБ был реализован проект по созданию стандартизированной и децентрализованной среды управления информацией электронных фондов Libmeta [11].. В проекте реализованы средства интеграции приложений с разными источниками/каталогами метаданных/данных, сервис директорий метаданных, унифицированный интерфейс поиска данных.

Существенное различие во внутренних моделях данных, используемых в различных музеях, библиотеках и архивах, является главной проблемой на пути решения задачи интеграции данных [10]. Для преодоления этой проблемы в решаемой задаче интеграции данных было предложено участникам экспортировать метаданные из своего внутреннего формата в формат на базе Dublin Core с использованием синтаксиса XML, так как во внутренних используемых форматах удается выделить общую часть, которая ложится в рамки предложенного формата. В системе создан универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH.

Основная коллекция метаданных была получена из библиотеки (тип источника внутренний) «Научное Наследие России» [12]. Для интеграции данных в LOD в качестве внешних источников было проведено связывание с данными DBpedia по авторам, а для связывания музейных экспонатов был проведен эксперимент с данными из Europeana.

Для каждого ресурса Libmeta может быть получено его представление, удовлетворяющее модели Europeana Semantic Elements (ESE) [15], которое определяет ряд обязательных элементов метаданных.

Для мониторинга новых данных и установления связей с внешними источниками данных в рамках системы используется SILK Framework [13]. Для установления связей необходимо указать источник данных, правила доступа к

данным и правила связывания. Вся эта информация была написана в виде конфигурационного файла на языке SILK LSL.

Сейчас проводятся работы по связыванию данных с авторитетными файлами VIAF[14] – это проект, который объединяет все значимые библиотеки, интегрирующие свои данные в LOD.

8. Заключение и дальнейшие работы

Разрабатываемая ПОЦБ предполагает поддержку функциональности, рекомендуемую проектом LOD, а именно: средства для представления информации из различных источников и для установления и поддержки связей между RDF-ресурсами, как внутренними, так и внешними, т.е. предполагает осуществление полного цикла интеграции набора данных в пространство LOD.

Основные преимущества реализации принципов LOD в Libmeta:

- Связность. Подключение источников, не обязательно библиотек;
- Машиночитаемость.Представление в RDF, использование общепринятых словарей и онтологий;
- Доступность. Доступные для свободного использования всеми пользователями без каких-либо ограничений в виде авторских прав.

Использование онтологии предметной области позволит не только включать другие типы ресурсов в библиотеку, но и уточнять и включать в библиотеку описания внутренней структуры информационных ресурсов нужной детализации, обращаясь за данными к источникам, которые раньше с трудом могли использоваться в рамках интеграции ресурсов электронных библиотек. В перспективе использование возможностей технологий и методов, предназначенных для извлечения информации из текстов, (называемые в англоязычных источниках термином text mining) для анализа сопутствующей текстовой информации и выявления неявных связей между различными объектами позволят решать задачи уточнения терминов онтологии предметной области и обработки текстовых документов для более точной их классификации.

Литература

1. Ле Хоай, А.Ф. Тузовский, Использование онтологии в электронных библиотеках, http://www.lib.tpu.ru/fulltext/v/Bulletin_TPU/2012/v320/i5/07.pdf, 2012,
2. О. М. Атаева, В. А. Серебряков, Подход к созданию персональной электронной семантической библиотеки, RCDL, 2013
3. Bernhard Haslhofer, Bernhard Schan, The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. http://eprints.cs.univie.ac.at/284/1/lodws2008.pdf, 2008

4. http://dbpedia.org
5. http://europeana.eu
6. Weber, R. Ontological Foundations of Information Systems , Queensland, Australia, Coopers & Lybrand. 1997.
7. О. М. Атаева, А. О. Еркимбаев, В. Ю. Зицерман, Г. А. Кобзев, К. П. Пушин, В. А. Серебряков, К. Б. Теймуразов. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования, RCDL, 2013
8. http://www.w3.org/TR/void/
9. Jain, P., Verma, K., Yeh, P.Z., Hitzler, P., Sheth, A.P.: LOQUS: Linked Open Data SPARQL Querying System. Technical report, Tech. rep., Kno. e. sis Center, Wright State University, Dayton, Ohio, 2010. Available from http://www. pascal-hitzler.de/resources/publications/loqus-tr-2010. pdf (2010)
10. А.Б. Антопольский, А.А. Каленкова, Н. Каленов, В.А. Серебряков, А. Сотников. Принципы разработки интегрированной системы для научных библиотек, архивов и музеев // Информационные ресурсы России. – 2012. - № 1. – С. 2-7.
11. А. Антопольский, О. Атаева, В. Серебряков. Среда интеграции данных научных библиотек, архивов и музеев «LibMeta» // Информационные Ресурсы России. – 2012. - №5.
12. http://e-heritage.ru/index.html
13. http://lod2.eu/Project/Silk.html
14. http://viaf.org/
15. http://pro.europeana.eu/ese-documentation/

LIBMETA

O.M. Atayeva, V.A. Serebryakov

The aim of this work is to develop an information system for creation of semantic digital library which content is individual for each user and is populated from various data sources located on the Web and integrated into LOD cloud. The system give the user a unified presentation in order to enable retrieval of the information on a certain subject area the user is interested in.